A Primer on PAC-Bayesian Learning

Benjamin Guedj

A PRIMER ON PAC-BAYESIAN LEARNING

by

Benjamin Guedj

Abstract. — Generalised Bayesian learning algorithms are increasingly popular in machine learning, due to their PAC generalisation properties and flexibility. The present paper aims at providing a self-contained survey on the resulting PAC-Bayes framework and some of its main theoretical and algorithmic developments.

Contents

2000 ***Mathematics Subject Classification.*** — 68Q32, 62C10, 60B10.
Key words and phrases. — PAC-Bayes, machine learning, statistical learning theory.

I would like to thank Agnès Desolneux and Mylène Maïda for giving me the opportunity to contribute to the 2nd SMF Congress in Lille, France. My deepest gratitude goes to colleagues, friends and co-authors who initiated or shared my unfailing interest in PAC-Bayes, with a special mention for Pierre Alquier, Olivier Catoni, Pascal Germain, Peter Grünwald, and John Shawe-Taylor.

1. Introduction

Artificial intelligence (AI) appears as the workhorse of a striking number of revolutions in several domains. As neurosciences, robotics, or ethics—to name but a few—shape new products, ways of living or trigger new digital rights, machine learning plays a more central role than ever in the rise of AI.

In the visionary words of Arthur Samuel (Samuel, 1959), machine learning is the field of study about computers' ability to learn without being explicitly programmed. As such, a long-term goal is to mimic the inductive functioning of the human brain, and most machine learning algorithms build up on statistical models to devise automatic procedures to infer general rules from data. This effort paved the way to a mathematical theory of learning, at the crossroads of computer science, optimisation and statistics (Shalev-Shwartz and Ben-David, 2014). The interest in machine learning has been considerably powered by the emergence of the so-called big data era (an abundance of data collected, and the alignement of the corresponding required computing resources), and attempts at unifying these research efforts have shaped the emerging field of *data science*.

Among several paradigms, the present paper focuses on a Bayesian perspective to machine learning. As in Bayesian statistics literature, Bayesian machine learning is a principled way of managing randomness and uncertainty in machine learning tasks. Bayes reasoning is all about the shift from inferring unknown deterministic quantities to studying distributions (of which the previous deterministic quantities are just an instance), and has proven increasingly powerful in a series of applications. We refer to the monograph Robert (2007) for a thorough introduction to Bayesian statistics.

Over the past years, several authors have investigated extensions of the celebrated Bayes paradigm. While these extensions no longer abide by the canonical Bayesian rules and may be harder to interpret by practitioners, they have been enjoying a growing popularity and interest from the machine learning community, where the focus is sometimes more on pure predictive performance than it is on estimation and explainability.

As an illustration, consider a supervised learning problem, with a regression instance: $Y = f(X) + W$ where $X \in \mathbb{R}^d$ (input), $Y \in \mathbb{R}$ (output) and $W \in \mathbb{R}$ (noise) are random variables. A typical Bayesian inference procedure for f (unknown – may be parametric, semiparametric or nonparametric) would focus on the posterior distribution given by

$$\text{posterior}(f|X,Y) \propto \text{likelihood}(X,Y|f) \times \text{prior}(f). \tag{1}$$

Note that when $f(X) = f_\theta(X) = \theta X$ (with $\theta \in \mathbb{R}^d$), one recovers the classical linear regression model (typically worked out under a Gaussian assumption for the noise W). To improve the model's flexibility and ability to capture a larger spectrum of

phenomena, it has been suggested by Zhang (2006a) to replace the likelihood by its *tempered* counterpart:

$$\text{target}(f|X,Y) \propto \text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(f), \tag{2}$$

where $\lambda \geq 0$ is a new parameter which controls the tradeoff between the *a priori* knowledge (given by the prior) and the data-driven term (the tempered likelihood). The resulting distribution (target) defines a different statistical modelling (possibly not in an explicit form). Note that (2) still defines a proper posterior: if $\lambda \leq 1$,

$$\begin{aligned}
&\text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(f) \\
&\leq \text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(f) \times \mathbb{1}[\text{likelihood}(X,Y|f) \geq 1] \\
&\quad + \text{prior}(f) \times \mathbb{1}[\text{likelihood}(X,Y|f) < 1] \\
&\leq \text{likelihood}(X,Y|f) \times \text{prior}(f) \times \mathbb{1}[\text{likelihood}(X,Y|f) \geq 1] \\
&\quad + \text{prior}(f) \times \mathbb{1}[\text{likelihood}(X,Y|f) < 1] \\
&\leq \text{likelihood}(X,Y|f) \times \text{prior}(f),
\end{aligned}$$

hence

$$\int \text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(\mathrm{d}f) \leq \int \text{likelihood}(X,Y|f) \times \text{prior}(\mathrm{d}f) + 1.$$

So the tempered posterior is proper as soon as the (non-tempered) posterior is. As for the case $\lambda \geq 1$,

$$\text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(f) \leq \left[\sup_{g \in \mathcal{F}} \text{likelihood}(X,Y|g)\right]^{\lambda} \times \text{prior}(f),$$

which yields, as soon as the likelihood is upper bounded (which is equivalent to assume that the MLE—maximum likelihood estimator—exists),

$$\int \text{likelihood}(X,Y|f)^{\lambda} \times \text{prior}(\mathrm{d}f) \leq \left[\sup_{g \in \mathcal{F}} \text{likelihood}(X,Y|g)\right]^{\lambda}$$

which makes the tempered posterior proper.

This tempered posterior notion[(1)] is at the core of the "safe Bayesian" paradigm (Grünwald, 2011, 2012, 2018; Grünwald and Van Ommen, 2017), where the parameter λ is integrated and marginalised out to yield more robust and automatic Bayesian inference procedures.

In machine learning, the emphasis on prediction ability is usually stronger than on inference (compared to the statistical literature). With that fact in mind, it is then only natural to go even further than the tempered likelihood: one can replace it by a purely arbitrary loss term, which only serves as a measure of the quality of prediction (*i.e.*, what loss is suffered when using the predictor g instead of f in the

[(1)] Interestingly, the multiplicative algorithm introduced by Vovk (1990) in online forecasting was later interpreted as an online version of such pseudo-posteriors.

previous example) and might not be supported by an explicit statistical modelling. This loss term is typically driven by information-theoretic arguments and therefore, substituting a loss term to the likelihood term achieves the shift from a model-based procedure to a purely data-driven procedure (which could arguably be described as *model-free*). Purely data-driven or model-free procedures may not assume an underlying probabilistic model to be inferred, but rather focus on an agnostic measure of performance.

In the sequel, we bundle under the term *generalised Bayes* such extensions towards tempered likelihoods or loss terms replacing likelihoods.

PAC-Bayesian inequalities were introduced by McAllester (1999a,b) based on earlier remarks by Shawe-Taylor and Williamson (1997). They have been further formalised by Seeger (2002), McAllester (2003a,b), Maurer (2004) and others. The goal was to produce PAC performance bounds (in the sense of a loss function) for Bayesian-flavored estimators – the term PAC-Bayes now refers to the theory delivering PAC bounds for *generalised* Bayesian algorithms (whether with a tempered likelihood or a loss term).

The acronym PAC stands for Probably Approximately Correct and may be traced back to Valiant (1984). A PAC inequality states that with an arbitrarily high probability (hence "probably"), the performance (as provided by a loss function) of a learning algorithm is upper-bounded by a term decaying to an optimal value as more data is collected (hence "approximately correct"). When applied to a Bayesian (or rather generalised Bayesian) learning algorithm, the theory is referred to as PAC-Bayesian. PAC-Bayes has proven over the past two decades to be a principled machinery to address learning problems in a striking variety of situations (sequential or batch learning, dependent or heavy-tailed data, etc.), and is now quickly re-emerging as a powerful and relevant toolbox to derive theoretical guarantees on the most recent learning topics, such as deep learning with neural networks or domain adaptation.

The rest of the paper is organised as follows. Section 2 introduces our notation, while Section 3 presents in more details generalised Bayesian learning methods. Section 4 contains a self-contained presentation of the PAC-Bayesian theory. Section 5 focuses on several practical implementations of PAC-Bayes and Section 6 illustrates the use of the PAC-Bayesian theory in several learning paradigms and some of its recent breakthroughs. Section 7 closes the paper.

2. Notation

The PAC-Bayesian theory has been successfully used in a variety of topics, including sequential learning (Gerchinovitz, 2011; Li et al., 2018), dependent or heavy-tailed data (Alquier and Guedj, 2018; Ralaivola et al., 2010; Seldin et al., 2012), classification (Lacasse et al., 2007; Langford and Shawe-Taylor, 2003; Parrado-Hernández

et al., 2012) and many others (see Section 6). To keep notation simple and still bear a fair amount of generality, we consider a simplified setting – let us stress however that results mentioned in this paper have been obtained in far more complex settings.

Let us assume that data comes in the form of a list of pairs $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ where each (X_i, Y_i) is a copy of some random variable $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ whose underlying distribution is denoted by $\mathbb{P}$. The goal is to build a functional object $\widehat{\phi}$ (which depends on $\mathcal{D}_n$) called a *predictor* such that for any new query X', $\widehat{\phi}(X') \approx Y'$ in a certain sense. In other words, learning is to be able to generalise to unseen data: this remark leads to *generalisation bounds*, also referred to as risk bounds, which are presented in Section 4. Note that predictors are functions $\mathbb{R}^d \to \mathbb{R}$; we call a *learning algorithm* a functional $\cup_{j=1}^{\infty}(\mathbb{R}^d \times \mathbb{R})^j \to \mathcal{F}$ which maps data samples to predictors (where $\mathcal{F}$ is the set of predictors). As such, we follow the notation used by Devroye et al. (1996) and focus on predictors in the sequel.

To assess the generalisation ability, we resort to a loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. Popular loss functions are the squared loss $\ell\colon (a,b) \mapsto (a-b)^2$, absolute loss $\ell\colon (a,b) \mapsto |a-b|$, 0-1 loss $\ell\colon (a,b) \mapsto \mathbb{1}[a \neq b]$, and so on. We then let

$$R\colon \widehat{\phi} \mapsto \mathbb{E}\left[\ell\left(\widehat{\phi}(X), Y\right)\right] \tag{3}$$

define the *risk* of the predictor $\widehat{\phi}$ (where the expectation is taken with respect to the underlying distribution of the data $\mathbb{P}$). As this underlying distribution is obviously unknown, the risk is not computable and is replaced by its empirical counterpart

$$r_n\colon \widehat{\phi} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{\phi}(X_i), Y_i\right). \tag{4}$$

As

$$\mathbb{E}\left[r_n\left(\widehat{\phi}\right)\right] = R\left(\widehat{\phi}\right),$$

we will see in Section 4 that obtaining PAC inequalities relies on how the process r_n concentrates to its mean R. Concentration inequalities such as Hoeffding's or Bernstein's are a key ingredient: we refer to the monograph Boucheron et al. (2013) for a thorough overview of concentration inequalities.

Let us now focus on the case where $\widehat{\phi}$ is a Bayesian predictor. The predictor $\widehat{\phi}$ may be of parametric, semiparametric, or nonparametric nature: in any case, a Bayesian approach would consider a prior distribution on such $\widehat{\phi}$: let us denote such a distribution π_0. Let us emphasise here that this prior operates on the collection of candidate predictors $\mathcal{F} = \left\{f\colon \mathbb{R}^d \to \mathbb{R}, f \text{ mesurable}\right\}$, or rather on a subspace $\mathcal{F}_0$ of it (*e.g.*, all linear functions from $\mathbb{R}^d$ to $\mathbb{R}$). A rich literature on model selection (either Bayesian or frequentist) studies refined inference techniques: see the monograph Massart (2007) for a solid introduction. One would materialise a statistical modelling

with a likelihood probability density function $\mathcal{L}$ and form the posterior distribution π of the model:

$$\pi\left(\widehat{\phi}|\mathcal{D}_n\right) \propto \mathcal{L}\left(\mathcal{D}_n|\widehat{\phi}\right) \times \pi_0\left(\widehat{\phi}\right). \tag{5}$$

Several inference techniques could then be derived from the posterior. For example, the mean of the posterior

$$\widehat{\phi}^{\text{mean}} = \mathbb{E}_\pi\, \phi = \int_{\mathcal{F}_0} \phi\pi(\mathrm{d}\phi),$$

its median

$$\widehat{\phi}^{\text{median}} = \text{median}(\pi),$$

the maximum a posteriori (MAP)

$$\widehat{\phi}^{\text{MAP}} \in \arg\max_{\phi\in\mathcal{F}_0} \pi(\phi),$$

or a single realisation

$$\widehat{\phi}^{\text{draw}} \sim \pi,$$

are all popular choices (with a slight abuse of notation, π refers to a probability measure or its density function, depending on context). The actual implementation of such predictors is discussed in Section 5. Theoretical results on Bayesian learning algorithms typically involve a thorough study of the way the posterior distribution concentrates as more data is collected. We refer the reader to the seminal papers Ghosal et al. (2000, iid case) and Ghosal and Van Der Vaart (2007, non-iid case). While Bayesian learning is a well established framework and is supported by theoretical and practical successes, a legitimate criticism is that its performance (both theoretical and practical) actually massively depends on the statistical modelling induced by the choice of the likelihood, the choice of the prior and possible hyperparameters, and any additional assumptions (such as an additive Gaussian noise, iid data, bounded functional, etc.). As famously stated by George Box[(2)], all modelling efforts form a subjective and constrained vision of the underlying phenomenon, which may prove herself of poor quality, if any. The past few decades have thus seen an increasing gap between the Bayesian statistical literature, and the machine learning community embracing the Bayesian paradigm – for which the Bayesian probabilistic model was too much of a constraint and had to be toned down in its influence over the learning mechanism. This movement gave rise to a series of works which laid down the extensions of Bayesian learning which are discussed in the next section.

[(2)] "Essentially, all models are wrong, but some are useful" (1976).

3. Generalised Bayesian learning

A first strategy consists in modulating the influence of the likelihood term, by considering a tempered version of it: from (5), the posterior now becomes the tempered posterior π_λ:

$$\pi_\lambda\left(\widehat{\phi}|\mathcal{D}_n\right) \propto \mathcal{L}\left(\mathcal{D}_n|\widehat{\phi}\right)^\lambda \times \pi_0\left(\widehat{\phi}\right), \tag{6}$$

where $\lambda \geq 0$. The former Bayesian model is now a particular case ($\lambda = 1$) of a continuum of distributions. Different values for λ will achieve different tradeoffs between the prior π_0 and the tempered likelihood $\mathcal{L}^\lambda$. Let us stress here that $\mathcal{L}^\lambda$ may no longer explicitly refer to a canonical probabilistic model.

This notion of tempered likelihood has been investigated, among others, by a striking series of paper (Grünwald, 2011, 2012, 2018; Grünwald and Van Ommen, 2017) which develop a "safe Bayesian" framework. These papers prove that the tempered posterior concentrates to the best approximation of the truth in the set of predictors $\mathcal{F}$, while this might not be the case for the non-tempered posterior: as such, tempering provides robustness guarantees when the chosen predictor, while being wrong, still captures some aspects of the truth.

We now rather focus on a second strategy which falls within generalised Bayes. Using an information-theoretic framework (see Csiszár and Shields, 2004, for an introduction) in which the "likelihood" of a predictor $\widehat{\phi}$ is no longer assessed by the probability mass from some specified model, but rather by the loss encountered when predicting $\widehat{\phi}(X)$ instead of Y, the actual output value we wish to predict.

In other words, the posterior from (5) or the tempered posterior from (6) are replaced with the *generalised posterior*

$$\pi_\lambda\left(\widehat{\phi}|\mathcal{D}_n\right) \propto \ell_{\lambda,n}\left(\widehat{\phi}\right) \times \pi_0\left(\widehat{\phi}\right), \tag{7}$$

where $\ell_{\lambda,n}$ is a loss term measuring the quality of the predictor $\widehat{\phi}$ on the collected data $\mathcal{D}_n$ (the training data, on which $\widehat{\phi}$ is built upon). To set ideas, one could think of $\ell_{\lambda,n}$ as a functional of the empirical risk r_n.

As the loss term is merely an instrument to guide oneself towards better performing algorithms but is no longer explicitly motivated by statistical modelling, the generalised Bayesian framework may be described as model-free, as no such assumption is required. Other terms appear in the statistical and machine learning literature, with occurrences of "generalised posterior", "pseudo-posterior" or "quasi-posterior" succeeding one another. Similarly, the terms "prior" and "posterior" have been consistently used as they "surcharge" the existing terms in Bayesian statistics, however the distributions in (7) are now different objects. Consider for example the prior π_0: rather than incorporating prior knowledge (which might not be available), π_0 serves as a way

to structure the set of predictors $\mathcal{F}_0$, by putting more mass towards predictors enjoying any other desirable property (suggested by the context, CPU / storage resources, etc.) such as sparsity.

From (7), the story goes on as in Bayesian learning: any mechanism yielding a predictor from the generalised posterior is admissible. As above, the mean, median, realisation or mode (MAP) are popular choices.

Among all possible loss functions $\ell_{\lambda,n}$, a most typical choice is the so-called Gibbs posterior (or measure):

$$\pi_\lambda\left(\widehat{\phi}|\mathcal{D}_n\right) \propto \exp\left[-\lambda r_n\left(\widehat{\phi}\right)\right] \times \pi_0\left(\widehat{\phi}\right). \tag{8}$$

The loss term exponentially penalises the performance of a predictor $\widehat{\phi}$ on the training data, and the parameter $\lambda \geq 0$ (often referred to as an inverse temperature, by analogy with the Boltzmann distribution in statistical mechanics) controls the tradeoff between the prior term and the loss term. Let us examine both extremes cases: when $\lambda = 0$, the loss term vanishes and the generalised posterior amounts to the prior: the predictor is blind to data. When $\lambda \to \infty$, the influence of data becomes overwhelming and the probability mass accumulates around the predictor[(3)] which achieves the best empirical error, *i.e.*, the generalised Bayesian predictor reduces to the celebrated empirical risk minimiser (ERM—see Vapnik, 1995, for a survey on statistical learning theory).

Why is the Gibbs measure so popular in machine learning? It arises in several contexts in statistics and statistical physics: let us illustrate this with a variational perspective. Let $(A, \mathcal{A})$ denote a measurable space and consider μ, ν two probability measures on $(A, \mathcal{A})$. We note $\mu \ll \nu$ when μ is absolutely continuous with respect to ν, and we let $\mathcal{M}_\nu(A, \mathcal{A})$ denote the space of probability measures on $(A, \mathcal{A})$ which are absolutely continuous with respect to ν:

$$\mathcal{M}_\nu(A, \mathcal{A}) = \{\mu \colon \mu \ \ll \nu\}.$$

We denote by $\mathcal{K}$ the Kullback-Leibler divergence between two probability measures:

$$\mathcal{K}(\mu, \nu) = \begin{cases} \int_{\mathcal{F}_0} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu & \text{when } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases} \tag{9}$$

Let us consider the optimisation problem

$$\underset{\mu \in \mathcal{M}_{\pi_0}(A,\mathcal{A})}{\arg\inf} \left\{ \int_{\mathcal{F}_0} r_n(\phi)\mu(\mathrm{d}\phi) + \frac{\mathcal{K}(\mu, \pi_0)}{\lambda} \right\}. \tag{10}$$

This problem amounts to minimising the integrated (with respect to any measure μ) empirical risk plus a divergence term between the generalised posterior and the

(3) There may be several predictors minimising the empirical risk.

prior. In other words, minimising a criterion of performance plus a divergence from the initial distribution, which is the analogous of penalised regression (such as Lasso). When $\mathcal{F}_0$ is finite and the loss is the squared loss $\ell\colon (a,b) \mapsto (a-b)^2$, one can easily deduce from the Karush-Kuhn-Tucker (KKT) conditions that the Gibbs measure π_λ in (8) is the only solution to the problem (10) (as proven by Rigollet and Tsybakov, 2012). In the general case, the proof is given by Lemma 2 in Section 4.

Let us also stress that the Gibbs posterior arises in other domains of statistics. Consider the case where the set of candidates $\mathcal{F}_0$ is finite. The mean of the Gibbs posterior is given by

$$\begin{aligned}\widehat{\phi}^{\text{mean}} := \mathbb{E}_{\pi_\lambda}\phi &= \int_{\mathcal{F}_0} \phi\, \pi_\lambda(\mathrm{d}\phi) \\ &= \int_{\mathcal{F}_0} \phi \exp\left[-\lambda r_n(\phi)\right] \pi_0(\mathrm{d}\phi) \\ &= \sum_{i=1}^{\#\mathcal{F}_0} \underbrace{\frac{\exp\left[-\lambda r_n(\phi_i)\right]\pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}_0}\exp\left[-\lambda r_n(\phi_j)\right]\pi(\phi_j)}}_{=:\,\omega_{\lambda,i}} \phi_i = \sum_{i=1}^{\#\mathcal{F}_0} \omega_{\lambda,i}\phi_i,\end{aligned}$$

which is the celebrated exponentially weighted aggregate (EWA, see for example Leung and Barron, 2006). EWA forms a convex weighted average of predictors, where each predictor has a weight which exponentially penalises its performance on the training data. Statistical aggregation (Nemirovski, 2000) may thus be revisited as a special case of generalised Bayesian posterior distributions (as studied in Guedj, 2013).

4. The PAC-Bayesian theory

The PAC learning framework has been initiated by Valiant (1984) and has been at the core of a great number of breakthroughs in statistical learning theory. In its simplest form, a PAC inequality states, for any predictor $\widehat{\phi}$ and any $\epsilon > 0$

$$\mathbb{P}\left[R\left(\widehat{\phi}\right) \leq \delta\right] \geq 1-\epsilon, \tag{11}$$

where δ is a threshold usually depending on data and ϵ. These risk bounds are of central importance in statistical learning theory as they give crucial guarantee on the performance of predictors, with an upper-bound and a confidence level ϵ which can be made arbitrarily small. When a matching lower bound is found, the predictor $\widehat{\phi}$ is said to be minimax optimal (see Tsybakov, 2003, and references therein). Note that in the original definition from Valiant (1984), the acronym PAC was used to refer to any bound valid with arbitrarily high probability together with the constraint that the predictor must be calculable in polynomial time with respect to n and $1/\epsilon$.

The acronym now has a broader meaning as it covers any risk bound holding with arbitrarily high probability.

PAC-Bayesian inequalities date back to Shawe-Taylor and Williamson (1997) and McAllester (1999a,b). McAllester's PAC-Bayesian bounds are empirical bounds, in the sense that the upper bound only depends on known computable quantities linked to the data.

***Theorem 1* (McAllester's bound).** — *For any measure $\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})$, and any $\epsilon > 0$,*

$$\mathbb{P}\left[\int R\left(\widehat{\phi}\right) \mathrm{d}\mu\left(\widehat{\phi}\right) \leq \int r_n\left(\widehat{\phi}\right) \mathrm{d}\mu\left(\widehat{\phi}\right) + \sqrt{\frac{\mathcal{K}(\mu, \pi_0) + \log \frac{2\sqrt{n}}{\epsilon}}{2n}}\right] \geq 1 - \epsilon. \tag{12}$$

The Kullback-Leibler term $\mathcal{K}(\mu, \pi_0)$ captures the complexity of the set of predictors $\mathcal{F}_0$. In the simplest case where $\mathcal{F}_0$ is a finite set of M predictors, this term basically reduces to M. If $\mathcal{F}_0$ is the set of linear functions, the complexity boils down to the intrinsic dimension d. More favorable regimes (of order $\log d$, under a sparsity assumption) have been obtained in the literature (see Guedj, 2013, for a survey). Overall, McAllester's bound expresses a tradeoff between empirical accuracy and complexity (in the sense of how far the posterior is from the prior).

This kind of bounds yields guarantees on the ("true") quality of the predictor $\widehat{\phi}$, with no need to evaluate or estimate its performance on some test data. This is a salient advantage of the PAC-Bayesian approach, as labelling and / or collecting test data might be cumbersome in some settings. Another key asset is that bounds of the form (12) are natural incentives to design new learning algorithms as minimisers of the right-hand side term (see Germain, 2015, for a discussion). By integrating out the whole expression over $\widehat{\phi}$, the constrained problem (10) appears once again and the Gibbs measure is deduced as the natural optimal generalised posterior distribution. McAllester's bounds have been improved by Seeger (2002, 2003) and Maurer (2004).

While of great practical use, McAllester's bounds did not hint about the rate of convergence of predictors, due to their empirical nature. Catoni (2004, 2007) therefore extended McAllester's PAC-Bayesian bounds to prove oracle-type inequalities, specifically on aggregated predictors (typically the mean of the Gibbs measure - see also Tsybakov, 2003, Yang, 2003, and Yang, 2004 for earlier works on aggregation and oracle inequalities in other settings than PAC-Bayes).

Catoni's technique consists of two ingredients:

1. A deviation inequality is used to upper bound the distance between $R\left(\widehat{\phi}\right)$ and its empirical counterpart $r_n\left(\widehat{\phi}\right)$ for a fixed $\widehat{\phi} \in \mathcal{F}_0$. In most of the rich PAC-Bayesian literature which followed Catoni's work, inequalities such as Bernstein's, Hoeffding's, Hoeffding-Azuma's or Bennett's have been used. More details can be found about these inequalities in the monographs Massart (2007) and Boucheron et al. (2013).
2. Then, the resulting bound is made valid for any $\widehat{\phi} \in \mathcal{F}_0$ simultaneously. Catoni suggests to consider the set of all probability distributions on $\mathcal{F}_0$ equipped with some suitable $\sigma-$algebra and make the deviation inequality uniform on this set with the following variational formula, presented in Lemma 1 (Legendre transform of the Kullback-Leibler divergence).

***Lemma 1* (Csiszár, 1975 ; Donsker and Varadhan, 1976 ; Catoni, 2004)**

Let $(A, \mathcal{A})$ be a measurable space. For any probability ν on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int(\exp \circ h)\mathrm{d}\nu < \infty$,

$$\log \int (\exp \circ h)\mathrm{d}\nu = \sup_{\mu \in \mathcal{M}_\nu(A,\mathcal{A})} \left\{ \int h\mathrm{d}\mu - \mathcal{K}(\mu, \nu) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of ν, the supremum with respect to μ on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\nu}(a) = \frac{\exp \circ h(a)}{\int(\exp \circ h)\mathrm{d}\nu}, \quad a \in A.$$

Proof. — Let $\mu \in \mathcal{M}_\nu(A, \mathcal{A})$. Since $\mathcal{K}(\cdot, \cdot)$ is non-negative, $\mu \mapsto -\mathcal{K}(\mu, g)$ reaches its supremum (equal to 0) for $\mu = g$. Then

$$\begin{aligned} -\mathcal{K}(\mu, g) &= -\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}g}\right)\mathrm{d}\mu \\ &= -\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu + \int \log\left(\frac{\mathrm{d}g}{\mathrm{d}\nu}\right)\mathrm{d}\mu \\ &= -\mathcal{K}(\mu, \nu) + \int h\mathrm{d}\mu - \log \int (\exp \circ h)\,\mathrm{d}\nu. \end{aligned}$$

Taking the supremum on all μ yields the desired result:

$$\log \int (\exp \circ h)\mathrm{d}\nu = \sup_{\mu \in \mathcal{M}_\nu(A,\mathcal{A})} \left\{ \int h\mathrm{d}\mu - \mathcal{K}(\mu, \nu) \right\},$$

□

Lemma 2. — *In Lemma 1, taking $\nu = \pi_0$ and $h = -\lambda r_n$ yields*

$$-\frac{1}{\lambda}\log \int \exp\left[-\lambda r_n(\phi)\right]\pi_0(\mathrm{d}\phi) = \inf_{\mu \in \mathcal{M}_{\pi_0}(A,\mathcal{A})} \left\{ \int r_n(\phi)\mu(\mathrm{d}\phi) + \frac{\mathcal{K}(\mu, \pi_0)}{\lambda} \right\}. \tag{13}$$

The unique distribution which achieves the minimum of the right-hand side is the Gibbs posterior given by (8)*, which solves problem* (10)*.*

Note that the second step in Catoni's technique requires to fix a reference measure ν on $\mathcal{F}_0$. The reference measure is used to control the complexity of set of predictors $\mathcal{F}_0$, however it kept being referred to as the "prior" to consistently extend the Bayesian setting (see Germain et al., 2016a, for a discussion on the links between Bayesian inference and PAC-Bayes). Catoni (2007) also makes connections with information theory and Rissanen's Minimum Description Length (MDL) principle (see Grünwald, 2007, for a solid introduction, and Zhang, 2006b, for the corresponding lower bounds). Other links have been studied between Catoni's bounds and generic chaining (Audibert and Bousquet, 2007) and fast rates (Audibert, 2009).

We can now state a general form for Catoni's bound (introduced in Catoni, 2003, 2004, 2007 and further extended by Audibert, 2004, 2010, Alquier, 2006, 2008, and Guedj, 2013, among others).

***Theorem 2* (Catoni, 2007).** — *Assume that the loss ℓ is upper bounded by some constant B. Consider the Gibbs measure defined in* (8)*. For any $\lambda > 0$, any $\epsilon > 0$,*

(14)
$$\mathbb{P}\left[\int R(\phi)\,\pi_\lambda(\mathrm{d}\phi) \leq \inf_{\mu\in\mathcal{M}_{\pi_0}(A,\mathcal{A})}\left\{\int R(\phi)\,\mu(\mathrm{d}\phi) + \frac{\lambda B}{n} + \frac{2}{\lambda}\left(\mathcal{K}(\mu,\pi_0) + \log\frac{2}{\epsilon}\right)\right\}\right] \geq 1-\epsilon.$$

As (14) holds for any $\lambda > 0$, we can now optimise the right-hand side to make the bound tighter, by using a union bound argument (as advised by Catoni, 2007, Sections 1.2 and 1.3, and Audibert, 2010, Section 2.2). The optimal value for λ in the right-hand side is given by

$$\lambda = \sqrt{\frac{2n\left[\mathcal{K}(\mu,\pi_0) + \log\frac{2}{\epsilon}\right]}{B}} \tag{15}$$

and denoting $\lambda^\star$ the optimal value in the left-hand side, and C a numerical constant, (14) becomes

(16)
$$\mathbb{P}\left[\int R(\phi)\,\pi_{\lambda^\star}(\mathrm{d}\phi) \leq \inf_{\mu\in\mathcal{M}_{\pi_0}(A,\mathcal{A})}\left\{\int R(\phi)\,\mu(\mathrm{d}\phi) + \sqrt{\frac{8B\left(\mathcal{K}(\mu,\pi_0) + \log\frac{2\log(nC)}{\epsilon}\right)}{n}}\right\}\right] \geq 1-\epsilon.$$

Note that this calibration of λ is purely theoretical and is useless in practice, as it depends on unknown terms: this is further discussed in Section 5.

Finally, note that assuming that the loss ℓ is convex yields a bound on the risk of the aggregated predictor $\widehat{\phi}^{\text{mean}}$ by a straightforward use of Jensen's inequality: (14) becomes

(17)
$$\mathbb{P}\left[R\left(\widehat{\phi}^{\text{mean}}\right) \leq \inf_{\mu \in \mathcal{M}_{\pi_0}(A,\mathcal{A})} \left\{ \int R\left(\phi\right) \mu\left(\mathrm{d}\phi\right) + \frac{\lambda B}{n} + \frac{2}{\lambda}\left(\mathcal{K}(\mu, \pi_0) + \log\frac{2}{\epsilon}\right)\right\}\right] \geq 1 - \epsilon.$$

Similar results have been obtained by Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Alquier and Biau (2013), Guedj and Alquier (2013) to derive PAC-Bayesian oracle inequalities for several sparse regression models. The key is to devise a prior π_0 which enforces sparsity, *i.e.*, gives larger mass to elements $\widehat{\phi} \in \mathcal{F}_0$ of small dimension (with respect to the sample size n).

The explicit tradeoff between accuracy and complexity brought by PAC-Bayesian bounds may be further controlled, with the notion of *localisation* (introduced by Catoni, 2004, formalised by Catoni, 2003 and further elaborated by Catoni, 2007, Section 1.3). Localisation consists in finely choosing the prior so as to reduce the Kullback-Leibler term. Two strategies have been investigated: data-dependent priors and distribution-dependent priors.

- As the prior cannot depend on the training data used to compute the empirical risk, one could split the initial data sample in two parts: one of them is then used to learn a relevant prior. This strategy has been applied by Ambroladze et al. (2007) and Germain et al. (2009), among others.
- Rather than depending on data, the prior can be made distribution-dependent, by directly upper-bounding the Kullback-Leibler divergence (Catoni, 2003, Ambroladze et al., 2007, Lever et al., 2010, 2013).

In particular, the localisation technique allows to remove the extra $\log(n)$ term in (16).

The PAC-Bayesian theory consists in producing PAC risk bounds (either empirical or oracle) of generalised Bayesian learning algorithms.

A slightly different line of work has also investigated similar results, holding in expectation rather than with high probability. While obviously weaker, such results have proven important in dealing with some settings (*e.g.*, with unbounded losses). Following a method initiated by Leung and Barron (2006), Dalalyan and Tsybakov (2007, 2008) replaced the first step in Catoni's technique (the deviation inequality) with Stein's formula. This technique was further investigated and improved in a series of papers (Dalalyan and Tsybakov, 2012a, Rigollet and Tsybakov, 2012, Alquier and Guedj, 2017).

Finally, let us mention that faster rates of convergence, of magnitude $\mathcal{O}(1/n)$, have been obtained by Audibert (2009); Audibert and Catoni (2011); Dinh et al. (2016); Grünwald and Mehta (2016); van Erven et al. (2015), to name but a few.

5. Algorithms: PAC-Bayes in the real world

In conclusion, the PAC-Bayesian framework enjoys strong theoretical guarantees in machine learning, in the form of (possibly minimax optimal) oracle generalisation bounds. However, the practical use of PAC-Bayes turns out to be a computational challenge when facing complex, high-dimensional data. As a matter of fact, PAC-Bayes faces the exact same issues as Bayesian learning, as in both cases one is often required to sample from a possibly complex distribution. In Bayesian learning, sampling from the posterior; in PAC-Bayes, sampling from the generalised posterior. Let us focus on the Gibbs posterior given by (8), as it is one of the most popular choices in PAC-Bayes. As in Bayesian learning, we often resort to a $d-$dimensional projection of the predictor $\widehat{\phi}$ or its development onto a functional basis (up to term K, for example). Monte Carlo Markov Chains (MCMC) are a popular choice for sampling from such a distribution. We refer to Andrieu et al. (2003) and Robert (2007) for an introduction to this (rich) topic, and to Bardenet et al. (2016) for a survey on most recent techniques for massive datasets.

The goal is to sample from the Gibbs measure

$$\pi_\lambda\left(\widehat{\phi}|\mathcal{D}_n\right) \propto \exp\left[-\lambda r_n\left(\widehat{\phi}\right)\right] \times \pi_0\left(\widehat{\phi}\right).$$

The analytical form of this distribution is known (as the prior π_0, the loss ℓ and the parameter λ are chosen). Three main techniques have been investigated in the literature.

1. The most popular one, by far, is MCMC. A naive pick is a Metropolis-Hastings algorithm (see Algorithm 1). However, due to the possibly high dimensionality of the generalised posterior π_λ, a nested model strategy coupled with a transdimensional MCMC is often a much better choice (as it could avoid sampling from a too high dimensional proposal distribution, for example). In that setting, the proposal distribution may yield states of different dimensions at each iteration. A simplified form of such a transdimensional algorithm (which was successfully applied to additive regression in Guedj and Alquier, 2013, binary ranking in Guedj and Robbiano, 2018 and online clustering in Li et al., 2018) is given by Algorithm 2. Other MCMC algorithms, such as Langevin Monte Carlo, have also been investigated (Dalalyan and Tsybakov, 2012b). MCMC algorithms (as in Algorithm 1 and Algorithm 2) output a sequence of points, whose stationary distribution is asymptotically the target p. Wether this property is reached for

Algorithm 1: Metropolis-Hastings algorithm

Input: Proposal q, target p, horizon T, initialisation x_0
Output: A sequence $(x_i)_{i=0}^T$

for $t = 1, \dots, T$ **do**
  $x \sim q$ // Sample a candidate state
  $\alpha := \min\left(1, \frac{p(x)}{p(x_{t-1})} \cdot \frac{q(x_{t-1})}{q(x)}\right)$ // acceptance ratio
  $U \sim \mathcal{B}(\alpha)$ // draw a Bernoulli trial
  if $U \equiv 1$ **then**
    $x_t := x$
  else
    $x_t := x_{t-1}$

a given number of iterations, or the quality of the approximation at a finite horizon, are central questions in the MCMC literature.

2. When using the mode of the Gibbs measure, *i.e.*,
$$\widehat{\phi}^{\text{mode}} \in \arg\sup_{\phi \in \mathcal{F}_0} \pi_\lambda = \arg\sup_{\phi \in \mathcal{F}_0} \left\{\exp\left[-\lambda r_n\left(\phi\right)\right] \pi_0\left(\phi\right)\right\},$$
it is often more efficient to resort to stochastic optimisation, such as gradient descent or its many variants (stochastic gradient descent, block gradient descent, to name a few). Gradient descent is one of the main workhorses of machine learning and we refer to Shalev-Shwartz and Ben-David (2014, Chapter 14) and references therein. A gradient-descent-based strategy has been applied in Alquier and Guedj (2017) for PAC-Bayesian-flavored non-negative matrix factorisation.
3. The third option which has been investigated in the literature is variational Bayes, which has gained a tremendous popularity in machine learning (see Wainwright and Jordan, 2008 ; also Blei et al., 2017, for a recent survey). It amounts to finding the best approximation of the Gibbs measure within a family of known measures, typically much easier to sample from. Alquier et al. (2016) propose an algorithm to find the best Gaussian approximation to the Gibbs measure (under assumptions on the prior and loss which make this approximation reasonably good).

Several works have contributed to bridging the gap between theory and implementations for PAC-Bayes.

1. For variational Bayes (Gaussian) approximation to the Gibbs measure, Alquier et al. (2016) show that whenever a PAC-Bayesian inequality holds for the Gibbs measure, a similar one (with the same rate of convergence) holds for the approximate generalised posterior (at the price of technical assumptions which control

Algorithm 2: A transdimensional MCMC algorithm adapted from Guedj and Alquier (2013)

Input: Family of proposals (q_j), target p, horizon T, initialisation x_0

`/* As many proposal distributions as nested models. A model is determined by which covariates from` $1, \dots, d$ `are selected. Two models sharing the same number of selected covariates are said to be neighbors. */`

Output: A sequence $(x_i)_{i=0}^T$

for *t=1,...,T* **do**
- Dimension shift: add one, remove one, or do nothing (each with probability 1/3)
- `neighbors` := set of models obtained from adding or substracting one unit to the dimension of the current model.
- **for** *each j in* ***neighbors*** **do**
  - $y_j \sim q_j$ `// e.g., a Gaussian`
- Pick model j with probability $\frac{p(y_j)/q_j(y_j)}{\sum_{k \in \text{neighbors}} p(y_k)/q_k(y_k)}$
- $\alpha := \min\left(1, \frac{p(y_j)}{p(x_{t-1}} \cdot \frac{q(x_{t-1})}{q_j(y_j)}\right)$ `// acceptance ratio`
- $U \sim \mathcal{B}(\alpha)$ `// draw a Bernoulli trial`
- **if** $U \equiv 1$ **then**
  - $x_t := y_j$
- **else**
  - $x_t := x_{t-1}$

the quality of the approximation in a Kullback-Leibler sense). This leads to a non-asymptotic control of the approximation error. This breakthrough allows for PAC-Bayesian oracle generalisation bounds on the actual algorithm which is implemented rather than on the theoretical object, and as such, echoes the celebrated statistical and computational tradeoff.

2. MCMC has been the most used sampling scheme in the PAC-Bayes literature, however very few results were available to guarantee its validity and quality in that setting. Li et al. (2018) proved that the stationary distribution is indeed the Gibbs measure for a particular model (online clustering). Note however that this is an asymptotic result: up to our knowledge, there is no non-asymptotic control of the approximation for Metropolis-Hastings-based algorithms. The Langevin Monte Carlo however, leads to a non-asymptotic control of the quality of the estimation error (see Dalalyan, 2017; Durmus et al., 2018).

As a concluding remark, let us examine how one should calibrate the parameter λ in practice. Two strategies are possible: cross-validation (yielding good results in practice, yet quite computationally demanding) and integration and marginalisation of λ similarly to what is proposed in the "safe Bayesian" framework (Grünwald, 2012).

6. Some recent breakthroughs in PAC-Bayes

Over the past years, the PAC-Bayesian approach has been applied to a large spectrum of settings. In addition to aforecited papers, let us mention classification (Germain et al., 2009), high-dimensional sparse regression (Alquier and Biau, 2013; Alquier and Lounici, 2011; Guedj and Alquier, 2013), image denoising (Salmon and Le Pennec, 2009), completion and factorisation of large random matrices (Alquier and Guedj, 2017; Mai and Alquier, 2015), recommendation systems, reinforcement learning and collaborative filtering (Ghavamzadeh et al., 2015), dependent or heavy-tailed data (Alquier and Guedj, 2018; Ralaivola et al., 2010; Seldin et al., 2012), co-clustering (Seldin and Tishby, 2010), meta-learning (Amit and Meir, 2018), binary ranking (Guedj and Robbiano, 2018; Li et al., 2013), transfer learning and domain adaptation (Germain et al., 2016b), online clustering (Li et al., 2018), algorithmic stability (London, 2017; London et al., 2014), multi-view learning (Sun, 2013; Zhao et al., 2017), variational inference in mixture models (Chérief-Abdellatif and Alquier, 2018), multiple testing (Blanchard and Fleuret, 2007), tailored density estimation (Higgs and Shawe-Taylor, 2010), etc.

A salient advantage of PAC-Bayes is its flexibility: the theory requires little assumptions to be applied to new topics and problems. The use of generalised Bayesian learning algorithms requires the definition of a loss, and of a prior (*i.e.*, a heuristics to navigate throughout the set of candidate predictors $\mathcal{F}_0$), which explains how it could have been applied to so many different learning settings.

Most recent works on PAC-Bayes have seen a growing interest in data-dependent priors (Dziugaite and Roy, 2018a,b) and distribution-dependent priors (Rivasplata et al., 2018). This movement can be seen as an additional layer of *generalisation*: since the model-based likelihood has been replaced by an agnostic data-driven loss term, why not shift from a model-constrained prior to a purely data-driven measure which captures elementary knowledge about the underlying phenomenon?

In the deep learning tide wave, the machine learning community (at large) has demonstrated the impressive empirical successes of neural networks in some tasks. However voices have risen to orient some of the research effort to obtain theoretical guarantees and bounds which would explain those successes. Very few results have been published, however a significant fraction of existing work massively relies on PAC-Bayes. Dziugaite and Roy (2017) and Neyshabur et al. (2017) prove generalisation bounds for neural networks, with computable bounds (inherited from

McAllester's initial bound) and numerical experiments proving the generalisation ability of (small) networks.

Last but not least, a few research efforts in the past years have focused on more agnostic and generic perspectives to obtain PAC-Bayes bounds, and to get rid of handy yet unrealistic assumptions such as boundedness of the loss function, or independence of data. Such assumptions allow for an extensive use of powerful mathematical results, and yet are hardly met in practice. Bégin et al. (2016) replaced the classical Kullback-Leibler divergence by the more general Rényi divergence, allowing to derive bounds in new settings. Alquier and Guedj (2018) then proposed an even more general divergence class, the f-divergences (of which the Rényi divergence is a special case).

PAC bounds for heavy-tailed random variables have been studied by Catoni (2004) under strong exponential moments assumptions. In a striking series of papers, several authors have taken over and improved those bounds with different tools: the small ball property (Grünwald and Mehta, 2016; Mendelson, 2015), robust loss functions (Catoni, 2016) and median-of-means tournaments (Devroye et al., 2016; Lugosi and Mendelson, 2016). However those papers mostly focus on linear regression (for predictors including the ERM, a minimiser of a modified loss function, or median-of-means–MoM). Alquier and Guedj (2018) derived PAC bounds with similar rates of convergence, holding for generalised Bayesian predictors. As for dependent data, several PAC or PAC-Bayesian bounds have been proven (Agarwal and Duchi, 2013; Ralaivola et al., 2010; Seldin et al., 2012) under boundedness or exponential moments assumptions.

Let us conclude this section by sketching the proof of the main result in Alquier and Guedj (2018). Note that data points are not required to be independent nor identically distributed. For the sake of concision we shall now omit the argument ϕ when no confusion can arise. We will use the notation $\psi_p(x) = x^p$.

Definition 1. — *For any $p \in \mathbb{N}$, let*

$$\begin{aligned}\mathcal{M}_{\psi_p,n} &:= \int_{\mathcal{F}_0} \mathbb{E}\left(\psi_p\left(|r_n(\phi) - R(\phi)|\right)\right) \pi_0(\mathrm{d}\phi) \\ &= \int_{\mathcal{F}_0} \mathbb{E}\left(|r_n(\phi) - R(\phi)|^p\right) \pi_0(\mathrm{d}\phi).\end{aligned}$$

Definition 2. — *Let f be a convex function with $f(1) = 0$. Csiszár's f-divergence between two measures μ and ν is given by*

$$D_f(\mu,\nu) = \begin{cases} \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{when } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that the Kullback-Leibler divergence in (9) is given by $\mathcal{K}(\mu,\nu) = D_{x\log(x)}(\mu,\nu)$.

Theorem 3 **(Alquier and Guedj, 2018).** — *Fix $p > 1$, $q = \frac{p}{p-1}$ and $\epsilon \in (0,1)$. With probability at least $1-\epsilon$ we have for any distribution μ*

$$\left| \int R \mathrm{d}\mu - \int r_n \mathrm{d}\mu \right| \leq \left(\frac{\mathcal{M}_{\psi_q,n}}{\epsilon} \right)^{\frac{1}{q}} \left(D_{\psi_p{-}1}(\mu, \pi_0) + 1 \right)^{\frac{1}{p}}.$$

Proof. — Let $\Delta_n(\phi) := |r_n(\phi) - R(\phi)|$. From Bégin et al. (2016), we derive:

$$\begin{aligned}
\left| \int R \mathrm{d}\mu - \int r_n \mathrm{d}\mu \right| &\leq \int \Delta_n \mathrm{d}\mu = \int \Delta_n \frac{\mathrm{d}\mu}{\mathrm{d}\pi_0} \mathrm{d}\pi_0 \\
&\leq \left(\int \Delta_n^q \mathrm{d}\pi_0 \right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi_0} \right)^p \mathrm{d}\pi_0 \right)^{\frac{1}{p}} \qquad \text{(Hölder ineq.)} \\
&\leq \left(\frac{\mathbb{E} \int \Delta_n^q \mathrm{d}\pi_0}{\epsilon} \right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi_0} \right)^p \mathrm{d}\pi_0 \right)^{\frac{1}{p}} \qquad \text{(Markov, w.p. } 1-\epsilon) \\
&= \left(\frac{\mathcal{M}_{\psi_q,n}}{\epsilon} \right)^{\frac{1}{q}} \left(D_{\psi_p{-}1}(\mu, \pi_0) + 1 \right)^{\frac{1}{p}}.
\end{aligned}$$

□

The message from Theorem 3 is that we can compare $\int r_n \mathrm{d}\mu$ (observable) to $\int R \mathrm{d}\mu$ (unknown, the objective) in terms of

- the moment $\mathcal{M}_{\psi_q,n}$ (which depends on the distribution of the data)
- the divergence $D_{\psi_p{-}1}(\mu, \pi_0)$ (which measures the complexity of the set $\mathcal{F}_0$).

As a straightforward consequence, we have with probability at least $1-\epsilon$, for any μ,

$$\int R \mathrm{d}\mu \leq \int r_n \mathrm{d}\mu + \left(\frac{\mathcal{M}_{\psi_q,n}}{\epsilon} \right)^{\frac{1}{q}} \left(D_{\psi_p{-}1}(\mu, \pi_0) + 1 \right)^{\frac{1}{p}},$$

which is a strong incitement to deduce the optimal generalised posterior as the minimiser of the right-hand side.

Definition 3. — *Define $\overline{r}_n = \overline{r}_n(\epsilon, p)$ as*

$$\overline{r}_n = \min \left\{ u \in \mathbb{R}, \int [u - r_n(\phi)]_+^q \, \pi_0(\mathrm{d}\phi) = \frac{\mathcal{M}_{\psi_q,n}}{\epsilon} \right\}.$$

The minimum always exists as the integral is a continuous function of u, is equal to 0 when $u = 0$ and $\to \infty$ when $u \to \infty$. We then define the optimal generalised posterior $\hat{\mu}_n$ as

$$\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\pi_0}(\phi) = \frac{[\overline{r}_n - r_n(\phi)]_+^{\frac{1}{p-1}}}{\int [\overline{r}_n - r_n(\psi)]_+^{\frac{1}{p-1}} \, \pi_0(\mathrm{d}\psi)}.$$

Alquier and Guedj (2018) then focus on the explicit computation of the two terms $\mathcal{M}_{\psi_q,n}$ and $D_{\psi_p{-}1}(\mu, \pi_0)$ in several cases: bounded and unbounded losses, iid or dependent observations, and prove the first PAC-Bayesian bound for a time series without any boundedness nor exponential moment assumption. As Theorem 3 is a

completely generic result and holds under no assumption whatsoever, it may serve as a starting point to derive existing PAC-Bayesian bounds (by adding assumptions).

7. Conclusion

As developed throughout the present paper, PAC-Bayesian learning is a flexible and powerful machinery, as it yields state-of-the-art oracle generalisation bounds under little assumptions for numerous learning problems.

A NIPS (now NeurIPS) 2017 workshop[(4)], an ICML 2019 tutorial[(5)] and the "PAC-Bayes" query on arXiv[(6)] illustrate how PAC-Bayes is quickly re-emerging as a principled theory to efficiently address modern machine learning topics, such as leaning with heavy-tailed and dependent data, or deep neural networks generalisation abilities.

Acknowledgements

The author is greatly indebted to an anonymous reviewer for providing insightful comments and constructive remarks which considerably helped improving the paper.

The author acknowledges financial support from Agence Nationale de la Recherche (ANR, grants ANR-18-CE40-0016-01-"BEAGLE" and ANR-18-CE23-0015-02-"APRIORI") and the Engineering and Physical Sciences Research Council (EPSRC, grant "MURI: Semantic Information Pursuit for Multimodal Data Analysis").

The author warmly thanks Omar Rivasplata for his careful reading and suggestions.

References

A. Agarwal and J. C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

P. Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6, 2006.

P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017. ISSN 1934-8045. doi: 10.3103/S1066530717010045. URL https://doi.org/10.3103/S1066530717010045.

(4) https://bguedj.github.io/nips2017/50shadesbayesian.html – slides and videos.
(5) https://bguedj.github.io/icml2019/index.html – slides and videos.
(6) https://arxiv.org/search/?query=PAC-Bayes&searchtype=all&source=header

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5690-0. URL https://doi.org/10.1007/s10994-017-5690-0.

P. Alquier and K. Lounici. PAC-Bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-taylor. Tighter PAC-Bayes bounds. In *Advances in neural information processing systems*, pages 9–16, 2007.

R. Amit and R. Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214, 2018.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

J.-Y. Audibert. *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis, Université Paris 6, 2004.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

J.-Y. Audibert. *Agrégation PAC-Bayésienne et bandits à plusieurs bras*. Habilitation à diriger des recherches, Université Paris-Est, 2010.

J.-Y. Audibert and O. Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(Apr):863–889, 2007.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 2016.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

G. Blanchard and F. Fleuret. Occam's hammer. In *International Conference on Computational Learning Theory*, pages 112–126. Springer, 2007.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

O. Catoni. A PAC-Bayesian approach to adaptive classification, 2003.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour 2001. Springer, 2004.

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics, 2007.

O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

B.-E. Chérief-Abdellatif and P. Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

I. Csiszár and P. C. Shields. *Information theory and statistics: a tutorial*. Now Publishers Inc, 2004.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory (COLT 2007), Lecture Notes in Computer Science*, pages 97–111, 2007.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012b.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 1996.

V. C. Dinh, L. S. Ho, B. Nguyen, and D. Nguyen. Fast learning rates with heavy-tailed losses. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 505–513. Curran Associates, Inc., 2016.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time — III. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.

A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2017.

G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *NeurIPS*, 2018a.

G. K. Dziugaite and D. M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pages 1376–1385, 2018b.

S. Gerchinovitz. *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*. PhD thesis, Université Paris-Sud, 2011.

P. Germain. *Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine*. PhD thesis, Université Laval, 2015.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML*, 2009.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016a.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. In *Proceedings of International Conference on Machine Learning*, volume 48, 2016b.

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5-6):359–483, 2015.

S. Ghosal and A. Van Der Vaart. Convergence rates of posterior distributions for non-iid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

S. Ghosal, J. K. Ghosh, and A. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

P. D. Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.

P. D. Grünwald. The Safe Bayesian. In *Algorithmic Learning Theory*, pages 169–183. Springer Berlin Heidelberg, 2012.

P. D. Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 195: 47–63, 2018.

P. D. Grünwald and N. A. Mehta. Fast Rates for General Unbounded Loss Functions: from ERM to Generalized Bayes. *arXiv preprint arXiv:1605.00252*, 2016.

P. D. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.

B. Guedj. *Aggregation of estimators and classifiers: theory and methods*. Theses, Université Pierre et Marie Curie - Paris VI, December 2013. URL https://tel.archives-ouvertes.fr/tel-00922353.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, 7:264–291, 2013. doi: 10.1214/13-EJS771. URL https://doi.org/10.1214/13-EJS771.

B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 196:70 – 86, 2018. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2017.10.010. URL http://www.sciencedirect.com/science/article/pii/S0378375817301945.

M. Higgs and J. Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *International Conference on Algorithmic Learning Theory*, pages 148–162. Springer, 2010.

A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Advances in Neural information processing systems*, pages 769–776, 2007.

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in neural information processing systems*, pages 439–446, 2003.

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *International Conference on Algorithmic Learning Theory*, pages 119–133. Springer, 2010.

G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

C. Li, W. Jiang, and M. Tanner. General oracle inequalities for Gibbs posterior with application to ranking. In *Conference on Learning Theory*, pages 512–521, 2013.

L. Li, B. Guedj, and S. Loustau. A quasi-Bayesian perspective to online clustering. *Electron. J. Statist.*, 12(2):3071–3113, 2018. doi: 10.1214/18-EJS1479. URL https://doi.org/10.1214/18-EJS1479.

B. London. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2931–2940, 2017.

B. London, B. Huang, B. Taskar, and L. Getoor. PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, pages 585–594, 2014.

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society (to appear)*, 2016.

T. T. Mai and P. Alquier. A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9 (1):823–841, 2015. doi: 10.1214/15-EJS1020. URL https://doi.org/10.1214/15-EJS1020.

P. Massart. *Concentration inequalities and model selection*. Springer, 2007.

A. Maurer. A note on the PAC-Bayesian Theorem. *arXiv preprint cs/0411099*, 2004.

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999b.

D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51 (1):5–21, 2003a.

D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003b.

S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):1–25, 2015. doi: 10.1145/2699439.

A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

B. Neyshabur, S. Bhojanapalli, D. A. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13: 3507–3531, 2012.

L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11(Jul):1927–1956, 2010.

P. Rigollet and A.B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 4(27):558–575, 2012.

O. Rivasplata, E. Parrado-Hernandez, J. Shawe-Taylor, S. Sun, and C. Szepesvari. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, pages 9214–9224, 2018.

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

J. Salmon and E. Le Pennec. An aggregator point of view on NL-Means. In *Wavelets XIII*, volume 7446, page 74461E. International Society for Optics and Photonics, 2009.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. ISSN 0018-8646. doi: 10.1147/rd.33.0210.

M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11:3595–3646, 2010.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. doi: 10.1145/267460.267466.

S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984.

T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16: 1793–1861, 2015. Special issue in Memory of Alexey Chervonenkis.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

V. G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory*, 1990.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, pages 783–809, 2003.

Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10 (1):25–47, 2004.

T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 10 2006a.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.

J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

Benjamin Guedj, Inria, Lille - Nord Europe research centre, France & University College London, Department of Computer Science and Centre for Artificial Intelligence, United Kingdom
E-mail : benjamin.guedj@inria.fr • *Url :* https://bguedj.github.io